Using Word Embeddings to Uncover Discourses

Quentin Dénigot and **Heather Burnett**
Laboratoire de Linguistique Formelle
5, rue Thomas Mann
F-75205 Paris Cedex 13
quentin.denigot@u-paris.fr
heather.susan.burnett@gmail.com

Abstract

Word embeddings are generally trained on very large corpora to ensure their reliability and to better perform in specific sets of tasks. Critical Discourse Analysis usually studies corpora of much more modest sizes, but could use the word similarity ratings that word embeddings can provide. In this paper, we explore the possiblity of using word embeddings on these smaller corpora and see how the results we obtain can be interpreted when synchronically analysing corpora from different groups.

1 Introduction

For the last few years, word embeddings have been used for a variety of tasks, from document classification (Kusner, Sun, Kolkin, & Weinberger, 2015) to sentiment analysis (Yu, Wang, Lai, & Zhang, 2017). Using the philosophical insights of the distributional hypothesis (Harris, 1954), they allow us to give an intuitive mathematical representation of words and their relationships to each other. They notably allow us to quantify the meaning differences of two words in the space using basic similarity metrics (cosine similarity), which has come with a number of interesting properties, including seemingly capturing non-trivial relationships between words, such as gender in some cases[1].

Some works have subsequently used these measures of similarity to assess semantic variation of words through time using large corpora (Garg, Schiebinger, Jurafsky, & Zou, 2018; Hamilton, Leskovec, & Jurafsky, 2016; Rudolph & Blei, 2018), seeing how certain abstract concepts have historically been associated with different groups, concepts or issues across time.

These methods, however, are typically used with extremely large corpora containing billions of tokens and therefore cannot be used in many endeavors in the humanities and social sciences, where the size of the corpora used is not only limited, but sometimes also impossible to augment in a sensible way[2]. There are many issues to having a corpus of a smaller size; one of them is the much discussed issue of *bias* in AI systems (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Garg et al., 2018; Gonen & Goldberg, 2019; Manzini, Lim, Tsvetkov, & Black, 2019; Zhao et al., 2019), whereby a corpus exhibiting a certain ideology that can easily be embedded in the relations between the word vectors will in fact lead to undesirable ideologically-laden similarity patterns. While that issue is not specific to smaller datasets and is present in all subfields of statistical learning, it may be more significant in smaller datasets. The other main issue is that, due to the stochastic nature of the algorithms used to construct word embeddings, running a given algorithm several times on the same corpus can lead to very different results (an issue henceforth referred to as the *stability* of the embeddings).

Both bias and stability can be serious issues for general-purpose language models, whose stated goals are higher performance in various tasks (such as those presented at SemEval). For the analysis of semantic variation across times or groups, however, the *bias* issue is to be seen as a feature rather than an issue: the language models generated then are to be used as tools to explore the corpora, if there are biases in word associations, they are among the things we want to keep intact.

Stability, on the other hand, remains a significant

[1] See the now famous example in (Mikolov, Chen, Corrado, & Dean, 2013) where *king* - *man* + *woman* = *queen*.

[2] For example, a study focusing on the works of one particular author cannot have a corpus that would be larger than the complete works of that author, which might turn out to be much smaller than the corpora generally used to create word embeddings.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 298-312.
Held on-line February 14-19, 2021

problem. Luckily, there have been attempts at solutions (Antoniak & Mimno, 2018). We argue here that once the *stability* issue has been acknowledged and is taken into account to mitigate the observed results in word embeddings, they can be a useful tool for Critical Discourse Analysis (CDA). In this work, we focus specifically on discourses about LGBT+ individuals in a political context, and the approach put forward here may help in uncovering how groups with opposing objectives with regards to LGBT+ rights view certain linguistic objects that then shape their discourse.

CDA involves "(a) finding a regular pattern in a particular text or set of texts [...] and then (b) proposing an interpretation of the pattern, an account of its meaning and ideological significance" (Cameron, 2001, p. 137). The step of identifying such regular patterns in discourses on LGBT+ people is usually done manually (Provencher, 2011; Van der Bom, Coffey-Glover, Jones, Mills, & Paterson, 2015; VanderStouwe, 2013). However, many researchers have found computational methods to be more systematic (Mautner, 2016). We argue embeddings generated by gay marriage debates can be used to identify discourses that previous methods miss, and illustrate this point using the 2013 debates on *le mariage pour tous* in the French Assemblée Nationale.

Computational research on LGBT+ legal rights in the UK has predominantly used keywords (Scott et al., 2001) to identify ideological differences in arguments for/against lowering the age of consent for gay sex (Baker, 2004), civil partnerships (Bachmann, 2011) and marriage (Findlay, 2017). *Keywords* are words used significantly more often in some texts rather than others, and they are discovered through comparing relative frequencies of supporters' and opponents' lexica to each other or to a reference corpus like the BNC. For each word in each corpus, their actual frequency is compared with an expected frequency computed using the two corpora's data. These values are then used to compute a *keyness factor* for each word in each corpus, typically using a χ^2 or log-likelihood metric. Ranking words by decreasing order of keyness in a corpus therefore allows us to see which words appear comparatively more in one corpus rather than in another, acting like "lexical signposts, revealing what producers of a text have chosen to focus on" (Baker, 2004, p. 90).

The keyword method is very useful in uncovering discourses in contexts where different groups will tackle different topics and therefore use a different lexicon. However, in cases where the groups in question use the same lexicon with a similar frequency, they cannot bring much information. Similarly they cannot be used to compare how a given word is used by different groups when there is no major difference in the number of times that word is uttered by either group. In other words, the keywords method can give us information on which words are favored in a given group compared with another, but they cannot tell us how these words are used, what they *mean* for the group in question.

For example, if we consider the French context, some discourses lack the "lexical signposts" that keywords allow us to uncover. In the case of the gay marriage debates that took place at the Assemblée Nationale between January 29th, 2013 and April 23rd, 2013, both sides of the debate (*pro* and *anti*) argued their views were in line with the Republican values of *liberté, égalité* and *fraternité* (1), (2).

(1) **Christiane Taubira** (Pro *mariage pour tous*) : Nous disons que le mariage ouvert aux couples de même sexe illustre bien la devise de la République. Il illustre la liberté de se choisir, la liberté de décider de vivre ensemble.
Yves Fromion (Against *mariage pour tous*) : Et la liberté des enfants d'avoir un père et une mère ?
Christiane Taubira: Nous proclamons par ce texte l'égalité de tous les couples, de toutes les familles.
Pierre Lequiller (Against mariage pour tous) : Et les enfants ?
Christiane Taubira : Enfin, nous disons aussi qu'il y a dans cet acte une démarche de fraternité, parce qu'aucune différence ne peut servir de prétexte à des discriminations d'État. January 29th, 2013

***C.T.**: We say that opening marriage to same-sex couples illustrates the Republic's motto. It illustrates the liberty of choice, the liberty of deciding to live together.*

***Y.F.**: What about the liberty for kids to have both a father and a mother?*

***C.T.**: With this text, we proclaim the equality of all couples and of all families.*

***P.L.**: What about the kids?*

***C.T**: Finally, we also say that this here is another*

step towards fraternity, because no difference can serve as an excuse for State discrimination.

(2) **Hervé Mariton** (Against mariage pour tous): Pour que la liberté soit aussi responsabilité, pour que l'égalité soit aussi respect de la différence, et pour que la fraternité se fonde, plutôt que sur la division, sur l'unité, nous ne voterons pas ce texte. January 29th, 2013

***H.M.**: For liberty to also be responsibility, for equality to also be the respect of difference, and for fraternity to be founded upon unity rather than division, we will not vote in favor of this text.*

While keywords approaches can help us determine the various themes that are tackled by one group or another on a given question, word embeddings allow us to specifically see how some words are used, and their semantic associations.

This paper will therefore focus on the following questions:

- Can word embeddings bring useful information for synchronic semantic analysis across groups?
- How can we address the issue of corpus size when using word embeddings?
- How can this tool be used to uncover discourses in French gay marriage debates and how does it compare to more traditional computational approaches (such as keywords)?

2 Corpus

The corpus consists of the debates surrounding the question of the legalisation of same-sex marriage at the French Assemblée Nationale. These discussions took place between January 29 and April 23 of the year 2013[3] and their transcript is freely available on the Assemblée Nationale's website[4]. There were 31 sessions discussing the issue at the Assemblée, the number of tokens for each corpus of interest is reported in table 1. The entire corpus was annotated to show the identity of the speaker for each utterance as well as their political group and whether they supported same-sex marriage or not (based on both their discourse and the final votes). The corpus is very asymmetrical in that the representatives that were against same-sex marriage spoke a lot more (despite being a minority), leading to the three corpora shown in Table 1: the "*all*" corpus, containing all utterances by all Representatives; the "*pro*" corpus, containing only the utterances of Representatives in favor of the adoption of the same-sex marriage; the "*anti*" corpus, containing only the utterances of Representatives against the adoption of same-sex marriage.

corpus	**all**	**pro**	**anti**
tokens	624 483	169 525	375 630

Table 1: This table shows the number of tokens per corpus. These are the numbers after the corpus has been cleaned from numerical characters.[5]

Our initial interest in this corpus stemmed from considerations of *dogwhistle politics* (Saul, 2018), which are situations where speakers will secretly signal their belonging to an ingroup while adressing a larger audience, sometimes simply by using terms in a way that is reminiscent of the sociolect of the ingroup in question. Regarding the issue of same-sex marriage itself, it was very contentious at the time of the debates in France. In particular, religious conservative groups were very outspoken against the law and led many of the popular uprisings against it that occurred (Béraud, Portier, Guyot-Sionnest, Wieviorka, & Ténédos, 2015; de Coorebyter, 2013). Due to historical, sociological, and political reasons (in particular due to the importance of the principle of *laïcité* 'secularism'), French conservatives might avoid using religious terminology and argumentation, especially at the Assemblée Nationale. Conservative politicians were therefore facing the conundrum of wanting to appeal to their religious voters without being in the full capacity of using religious speech. We therefore hypothesize that they might use some secular words (like *nature* or *civilisation*) which, in their

[3] This includes a pause in the debates between February 12 and April 17, when a new version of the text was being written.

[4] http://www.assemblee-nationale.fr/14/dossiers/mariage_personnes_meme_sexe.asp

[5] The reason why the sum of tokens for the *pro* and the *anti* corpora does not add up to the number of tokens for the all corpus is that the many utterances produced by the presiding representative are omitted in the two position-specific corpora. The president's role in these debates is mostly to announce votes and give the floor to the next speaker, but they do not take part in the debates while they are on president duty, meaning their highly normalized discourse cannot be clearly defined as being *pro* or *anti* (although they do get to vote in the end).

mouths and in the context of this debate, would acquire religious connotations.

Likewise, one recurring complaint against conservative politicians at the time was that the use of the slippery slope argument according to which legalising same-sex marriage would lead to *PMA* ('In Vitro Fertilization') being available to more couples, which would in turn lead to *GPA* ('surrogacy'), which is illegal in France, to also become legal:

(3) **Philippe Meunier** (Against *mariage pour tous*) : Aujourd'hui le mariage, demain la PMA, et nous savons qu'au sein de la majorité certains souhaitent la GPA. February 2nd, 2013
Marie-Georges Buffet (Pro *mariage pour tous*) : Par ailleurs, chers collègues de l'opposition, vous ne cessez de vouloir lier la PMA et la GPA, au nom de l'égalité.
(...)
Annie Genevard (Against *mariage pour tous*) : Nous pensons que la PMA constitue, avec la GPA, le véritable objectif des partisans du projet de loi. February 3rd 2013

***P.M.**: Today marriage, tomorrow IVF, and we know that some among the majority wish for surrogacy.*

***M.-G.B.**: By the way, dear colleagues from the opposition, you cannot refrain from linking IVF to surrogacy in the name of equality.*

***A.G.**: We think that IVF, along with surrogacy, constitutes the actual goal of the supporters of this text.*

We hypothesize that terms like *PMA*, *GPA* and *mariage* should be seen as related to each other. Because of other concerns at the time that politicians subtly presented the legalisation of same-sex marriage as equivalent to the legalisation of pedophilia, polygamy, zoophilia and other crimes and infractions[6], we hypothesize that terms such as *PMA* and *GPA*, and possibly even *mariage*, could also be associated with these other, more taboo, illegal practices.

Overall, the resulting situation is one in which various groups of listeners will have various interpretations of a single message based on their knowledge about the political orientation of the speaker. It is therefore possible that a given word is used with different intended meanings according to the speaker's group. We hypothesize that such different meanings lead to different uses of the word and different collocations, following the basic ideas behind distributional semantics models and are therefore likely to lead to variation between our corpora.

3 Methods

Distributional semantics (Harris, 1954) is a branch of semantics that focuses on the idea that the meaning of a word can be derived from the words that appear in the same context (known as *distributional hypothesis*). The word "context" in this case is usually defined syntactically as the words immediately preceding and following the word of interest. Words that have similar contexts will therefore be characterized as being more similar overall, either through simple collocation or through shared collocates.

For example, one might want to know which words are similar to the word "salt" and might observe that the word "pepper" not only appears very often alongside "salt" (i.e. is a collocate) but also, like "salt", appears very often alongside words like "seasoning" or "spice" (i.e. shares collocates with "salt"). This second point is interesting with regards to applying the distributional hypothesis, since two words of identical (or nearly-identical) meaning are in fact unlikely to appear together, but are likely to appear alongside the same words and therefore share their distribution. This approach can lead to a formal definition of synonymy in terms of the similarity of the syntactic context for two words in a given pair: imagining each word as a vector in a vector space, with each dimension of each vector containing the count of how many times this word appears with a specific different word, we can then approximate how similar two distributions are (e.g.: using the dot-product between those two vectors). This allows us to treat each word as a point in a multidimensional space: for any word w, words similar (in a distributional sense) to w will occupy points in space that are close to that of w, while words which are dissimilar to w will occupy other parts of the space. Once we have this vector space at hand, different operations can be performed, including finding which words are viewed as more similar

[6] https://www.sos-homophobie.org/mariage-pour-tous-et-toutes/charte

to w in the model. These words will be those that optimize a given measure of vector distance, like cosine similarity.

In the context of this work, we have used the word2vec approach, first presented in Mikolov et al. (2013). Word2vec is a shallow, two-layer neural network which takes as input a large corpus and outputs a vector-space of several hundred dimensions in which each unique word of the corpus is assigned a vector. In this work, we specifically used the CBOW (Continuous Bag-of-Words) algorithm, which works as described above, by trying to predict a target word from its surrounding context[7].

Word2vec and similar algorithms have notably been used to track the meaning associations of words over time with interesting results (Garg et al., 2018). In our case, we would like to compare word meaning associations not over time, but across ideologically-opposed communities. The task is similar: we have a corpus that we split into smaller corpora depending on political orientation. An issue that we face, which does not apply to Garg et al. (2018) and others, is corpus size. The corpora used in the works cited above are large corpora, and in fact the corpora used to train algorithms like word2vec are typically a lot larger than the corpus we have here (see Table 1). This can lead to several issues: first, it is possible that the word similarity results that we get are not very precise, especially for less frequent words; second, because word2vec is initiated using random weights, two iterations of the algorithm do not necessarily produce identical results. Because the corpus is smaller, it can be the case that these results differ immensely.

This is not ideal. Luckily, there are solutions to mitigate these issues and ensure the stability of our results (or at least measure how certain we can be about the output). These solutions are notably described in Antoniak and Mimno (2018), and then applied in Rodman (2020).

A first solution to these issues relies on *fine-tuning* (Howard & Ruder, 2018), which is the practice of taking a vector space model trained on a larger corpus and retraining it on a new corpus for specific purposes. It is assumed in these cases that the larger corpus is large enough for the smaller corpus' vocabulary to be a subset. The model trained on the larger corpus (the *pre-trained* model) is reliable in its predictions regarding word similarities. By retraining it on our corpus, we bias it towards the distribution that our corpus has, allowing us to profit from the larger vocabulary size and stability of the pre-trained model while having results pertaining to our own corpora.

Even with fine-tuning however, there remain a number of things that can influence the configuration of our final vector space. As stated in Antoniak and Mimno (2018), the presence of specific documents in a corpus can have significant effects on the cosine similarities between embedding vectors. Luckily, Antoniak and Mimno (2018) also show that one can produce reliable outputs from smaller corpora by controlling for the presence of specific documents and their lengths through bootstrapping of the corpora. The corpus under study is bootstrapped at the document scale and new models are generated for each bootstrapped version of the corpus; the diverging model outputs can then be averaged, leading to stable results even for smaller corpora.

Following Rodman (2020), we decided to both fine-tune the model to our corpora and to bootstrap them and generate several models, the outputs of which we then averaged to obtain the results presented in section 4. The pre-trained model is one of those found in (Fauconnier, 2015); it outputs 500-dimensions vectors and was trained on a lemmatized version of the frWac corpus. The frWac corpus (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009) is a 1.3 billion word web-crawled French corpus. Because of the origin of the documents it contains (various websites of the .fr domain), it is likely to contain language that is different from the language that can be found in our own corpora. It was also constructed a few years before these debates were at the center of attention in France. Nevertheless, because of its size, it gives a solid base for the meaning of most common words found in our own corpora.

For the bootstrapping phase, the question arose of choosing which units in the corpus we wanted to bootstrap. We used *utterances* as our basic units. The reasoning behind this is that each intervention by a representative at the Assemblée is supposed to be self-contained (the speeches are prepared in advance), whereas the sessions themselves contain many utterances each and are a lot more variable.

[7]The other algorithm, *skip-gram*, predicts the surrounding context given a target word. Both algorithms could be used here, we preferred CBOW because it is computationally lighter than skip-gram. It tends, however, to smooth the context and although it gives accurate predictions for more frequent words, less frequent words' results are more erratic than they would be using skip-gram.

Utterances were defined as *uninterrupted* interventions by one or more representatives. Most of the interruptions are due to heckling from the audience. Because such interruptions sometimes force the speakers to address their audience before resuming with a linguistic blank slate, the two parts of a discontinuous intervention were treated as separate utterances. Once we have generated a model from each of the bootstrapped versions of our data, we generate the lists of closest semantic neighbours by computing the mean cosine similarity between words across all the models generated, along with its confidence interval (following Antoniak and Mimno 2018).

A final issue that remains is what Rodman (2020) calls "spatial noncomparability". The issue here is that the fine-tuning step in our work can lead to radical changes in the shape of our vector space. In short: we are not immune to the fact that the training algorithm might change the space in different ways according to its initial weights and to the corpus it is working on. In our case, the pre-training ensures that we start with the same initial weights, however, given that our corpora and their bootstrapped versions are rather small, it is possible that the presence of some utterances drastically alter the general layout of the vector space (especially the longer utterances).

This has one key consequence: the cosine similarity scores that we derive for the models are not necessarily directly comparable across corpora. This has to be kept in mind when analyzing the outputs of the models. Thanks to the fine-tuning and averaging processes that our outputs undergo, the risk that the outputs are completely incomparable is mitigated, but it is still present. This is why, even though we will display the average cosine similarity scores, we do not consider them to be as important as the ordering of word similarities (which should not be impacted by spatial noncomparability).

As an additional sanity check, we first focused on words for which we do not expect the semantic neighbours across corpora to be very different. The words we chose for this test were words which we assumed were used similarly by both sides of the debate. These words include deictic words (*ici, hier...* here, yesterday...) as well as words specific to the legal sociolect used at the Assemblée (*séance, amendement, article...* session, amendment, article...). The full list of sanity check words is presented in Table 2.

While some difference is to be expected due to the stochastic nature of the process, when we compare the closest semantic neighbours to these words in the *pro* and then the *anti* corpus, this difference should not be as great as with words we suspect are used differently. We assess the differences by counting how many words among the 12 closest semantic neighbours are identical across models. This initial count of differences is available in the central column in Tables 2 and 3.

The words we investigated expecting to find interesting differences can be found in Table 3, again, with the count of identical closest semantic neighbours across models. These words were chosen following intuitions that we had from reading the corpus itself. As mentioned earlier, the values of "*liberté*", "*égalité*", and "*fraternité*" are all invoked to defend different positions. The words "*nature*" and "*naturel*" were also tested, as their use has been commented on in Théry and Portier (2015), seemingly indicating a rupture between secular discourses and Catholic conservative discourses . "*Mariage*" and "*famille*" were also used, given that the debates at their core questioned their definitions. We can also add the word "*adoption*", because adoption was also extended to same-sex couples with this text. Finally, "*GPA*" and "*PMA*" were studied because of the slippery slope argument put forward by the opposition (see (3)).

All the words presented in Tables 2 and 3 were already present in the vocabulary of the pre-trained model.

4 Results

We can see from Tables 2 and 3 that a simple numerical analysis is not as eloquent as we thought it would be. We assumed that our control words would not vary much across corpora and that our test words might vary more. This is not backed by our data, there does not appear to be a difference threshold that allows us to fully differentiate between the two sets of words.

Because that initial measurement is rather weak and only concerns the 12 closest semantic neighbours for each word, we have tried to find a measure to calculate the difference at the scale of the entire vocabulary of the model. In order to do so, we have taken the intersection of the vocabularies of the two model families (*pro* and *anti*), we then used the ordering of similarity for all words in the *pro* models, with each word a label associated with

word	difference	ρ
amendement	7	0.02754716
article	7	0.001340266
assemblée	3	0.01551367
député	3	0.0954906
hier	6	0.1340674
ici	7	0.06237997
loi	6	0.01623462
président	7	0.05851079
rapporteur	3	-0.01754226
vote	1	0.04566755

Table 2: These are the control words used in our study along with the difference measurements between *pro* and *anti* models. The central column is the initial count-based difference measurement between lists of closest semantic neighbours. 12 is the biggest possible score (meaning all 12 closest semantic neighbours are different between *pro* and *anti*). The rightmost column is the Spearman's ρ score that was computed when attempting to do the more in-depth comparison mentioned in section 4.

word	difference	ρ
mariage	7	0.04984883
nature	5	0.0416876
naturel	4	0.08985369
famille	8	0.05748241
GPA	8	0.05427596
PMA	11	0.06521176
liberté	4	0.1114692
égalité	3	0.02781825
fraternité	1	0.1005997
adoption	9	0.007430747

Table 3: These are the test words used in our study along with the difference measurements between *pro* and *anti* models. The column layout is the same as that in Table 2

a position in the list, and compared it with the ordering of words in the *anti* models. Using the same label-number pairing as the *pro* models, we were able to end up with two ordered lists of numbers, one for each model family. We have then used Spearman's ρ to check how different the ordering of the two lists were. The results are displayed in the rightmost column of Tables 2 and 3. These measurements do not appear to show a clear distinction between control words and test words either, it turns out that the fine-tuning of the model brings along a great number of changes also in the least similar words.

That being said, we can still look in more detail at what the differences actually are. Figures 1 and 2 can show us a clearer picture by comparing these lists of closest semantic neighbours directly. If we compare two words that obtained similar difference scores, like *mariage* for the test words and *amendement* for the control words, we can see that although they are rated similarly, the differences they display are not necessarily similar. In the case of *amendement*, the closest semantic neighbours across corpora are not the same, but they mostly belong to the same subset of specialized lexicon. This is also what is observed for the rest of the control words[8]. Regarding *mariage*, however, the differences are more interesting in terms of discourse. For example, the word *mariage* in the *anti* corpus has among its closest semantic neighbours the word *filiation*, in keeping with the conservative idea that marriage is conceived first and foremost through the lens of procreation.

The word *adoption* has *homoparentalité* ('homoparentality') as closest semantic neighbour in the *anti* corpus, and indeed adoption by same-sex couples is a key issue in the *anti* discourse. The word *famille* is more similar to *société* than it is to *familial* in the *anti* corpus, whereas the word *société* is not among the 12 closest semantic neighbours in the *pro* corpus, underlying again the conservative ideal of society being built on the traditional family.

Regarding *PMA* and *GPA*, we can observe the interesting fact that while *PMA* appears as one of the closest semantic neighbours to *GPA* in both corpora, that relation is not symmetrical. The word *GPA* is not among the closest semantic neighbours to *PMA* in the *pro* corpus, whereas it is the third

[8]With the notable exceptions of *rapporteur*, where we find a number of proper nouns, and *ici*, which is mostly associated with website interfaces.

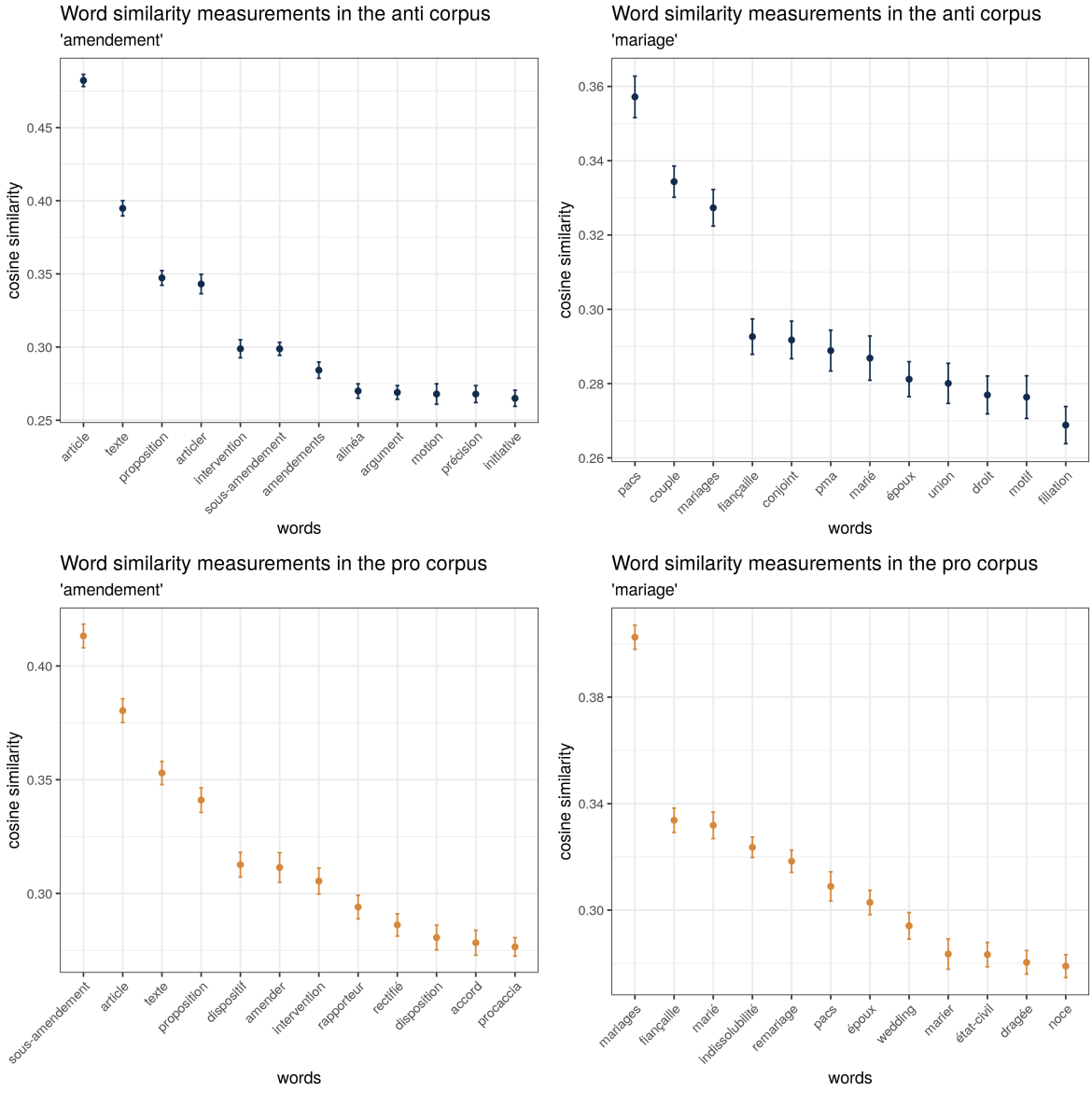

Figure 1: Comparison of closest semantic neighbours for *amendement* across models. The error bars are 95% confidence intervals, showing that our methods have led to reasonably stable similarity orderings. y axes show the mean cosine similarity across models between *amendement* and the words on the x axes. The models trained on the *anti* corpus are on top, the ones trained on the *pro* corpus are on the bottom.

Figure 2: Comparison of closest semantic neighbours for *mariage* across models. The error bars are 95% confidence intervals, showing that our methods have led to reasonably stable similarity orderings. y axes show the mean cosine similarity across models between *mariage* and the words on the x axes. The models trained on the *anti* corpus are on top, the ones trained on the *pro* corpus are on the bottom.

closest semantic neighbour to *PMA* in the *anti* corpus, in keeping with the confusion that was maintained by conservatives between the two. See Appendix A for this comparison and a few others.

5 Discussion

We have shown that from a theoretical standpoint, word embeddings can give us information that more traditional computational approaches to discourse analysis cannot. Indeed, whereas approaches like *keywords* give us information about *which* words are used more by one group compared with another, word embeddings can potentially tell us *how* a given word is used by a given group.

Using the existing literature, we see that it is possible to ensure the stability of word vectors trained on smaller datasets, or at least to have a reasonable estimation of the uncertainty of the results.

Because of their nature, these tools can expand the horizons of what can be considered to be exploitable text data in CDA, allowing to have distributional information on corpora that would be too big to study by hand, even though they would be too small to be used as training data to perform well on traditional NLP tasks.

While the purely quantitative approach does not allow us to isolate the interesting differences in distribution in our corpora, we have seen that these techniques can still allow us to explore the data in new ways that can complement a qualitative review of discourses. We are positive that with some refinement, these tools combined with human investigation of the texts could produce interesting analyses of socially marked corpora and be used for the synchronic comparison of discourses.

6 Conclusion

This paper focused on the use of existing computational methods applied to the quantitative analysis of modest size corpora. While there is already work using these methods to document semantic shifts across time, including some using corpora of a size similar to ours, this approach is not generally used to study semantic shifts across communities. Ultimately, the techniques applied to circumvent the issue of corpus size did not lead to the stability expected to conduct systematic quantitative analysis of the data, but the results obtained still lend themselves to interesting qualitative analyses going beyond the kind of conclusions that can be reached using more standard approaches to discourse analysis.

7 Acknowledgements

We would like to thank Jamie Findlay, Denis Paperno and Théis Bazin for their insightful contributions and feedback. This research was supported in part by the European Research Council (ERC) grant *Formal Models of Social Meaning and Identity Construction* (grant agreement N° 850539) as well as by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001

References

Antoniak, M., & Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119.

Bachmann, I. (2011). Civil partnership–"gay marriage in all but name": A corpus-driven analysis of discourses of same-sex relationships in the uk parliament. *Corpora*, *6*(1), 77–105.

Baker, P. (2004). 'unnatural acts': Discourses of homosexuality within the house of lords debates on gay male law reform. *Journal of sociolinguistics*, *8*(1), 88–106.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209–226.

Béraud, C., Portier, P., Guyot-Sionnest, P., Wieviorka, M., & Ténédos, J. (2015). *Métamorphoses catholiques: Acteurs, enjeux et mobilisations depuis le mariage pour tous*. Éditions de la Maison des sciences de l'homme, Paris. Retrieved from https://books.google.fr/books?id=5tnxDwAAQBAJ

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349–4357).

Cameron, D. (2001). *Working with spoken discourse*. Sage.

de Coorebyter, V. (2013). Le retour de la vieille france. *Le Soir*. Retrieved 2020-10-08, from http://www.crisp.be/2013/04/retour-vieille-france/

Fauconnier, J.-P. (2015). *French word embeddings*. Retrieved from http://fauconnier.github.io

Findlay, J. Y. (2017). Unnatural acts lead to unconsummated marriages: Discourses of homosexuality within the house of lords debate on same-sex marriage. *Journal of Language and Sexuality*, *6*(1), 30–60.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning* (pp. 957–966).

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to cda. *Methods of critical discourse studies*, *3*, 155–180.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Provencher, D. M. (2011). 'i dislike politicians and homosexuals': Language and homophobia in contemporary france. *Gender & Language*, *4*(2).

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, *28*(1), 87–111.

Rudolph, M., & Blei, D. (2018). Dynamic embeddings for language evolution. In *Proceedings of the 2018 world wide web conference* (pp. 1003–1011).

Saul, J. (2018). Dogwhistles, political manipulation, and philosophy of language. *New work on speech acts*, *360*, 84.

Scott, M., et al. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the wordsmith tools suite of computer programs. *Small corpus studies and ELT*, 47–67.

Théry, I., & Portier, P. (2015). Du mariage civil au mariage pour tous. sécularisation du droit et mobilisations catholiques. *Sociologie (online)*, *6*(1). Retrieved from https://journals.openedition.org/sociologie/2528

Van der Bom, I., Coffey-Glover, L., Jones, L., Mills, S., & Paterson, L. L. (2015). Implicit homophobic argument structure: Equal-marriage discourse in the moral maze. *Journal of Language and Sexuality*, *4*(1), 102–137.

VanderStouwe, C. (2013). Religious victimization as social empowerment in discrimination narratives from california's proposition 8 campaign. *Journal of Language and Sexuality*, *2*(2), 235–261.

Yu, L.-C., Wang, J., Lai, K. R., & Zhang, X. (2017, September). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 534–539). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D17-1056 doi: 10.18653/v1/D17-1056

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

A More interesting figures

A.1 Test words

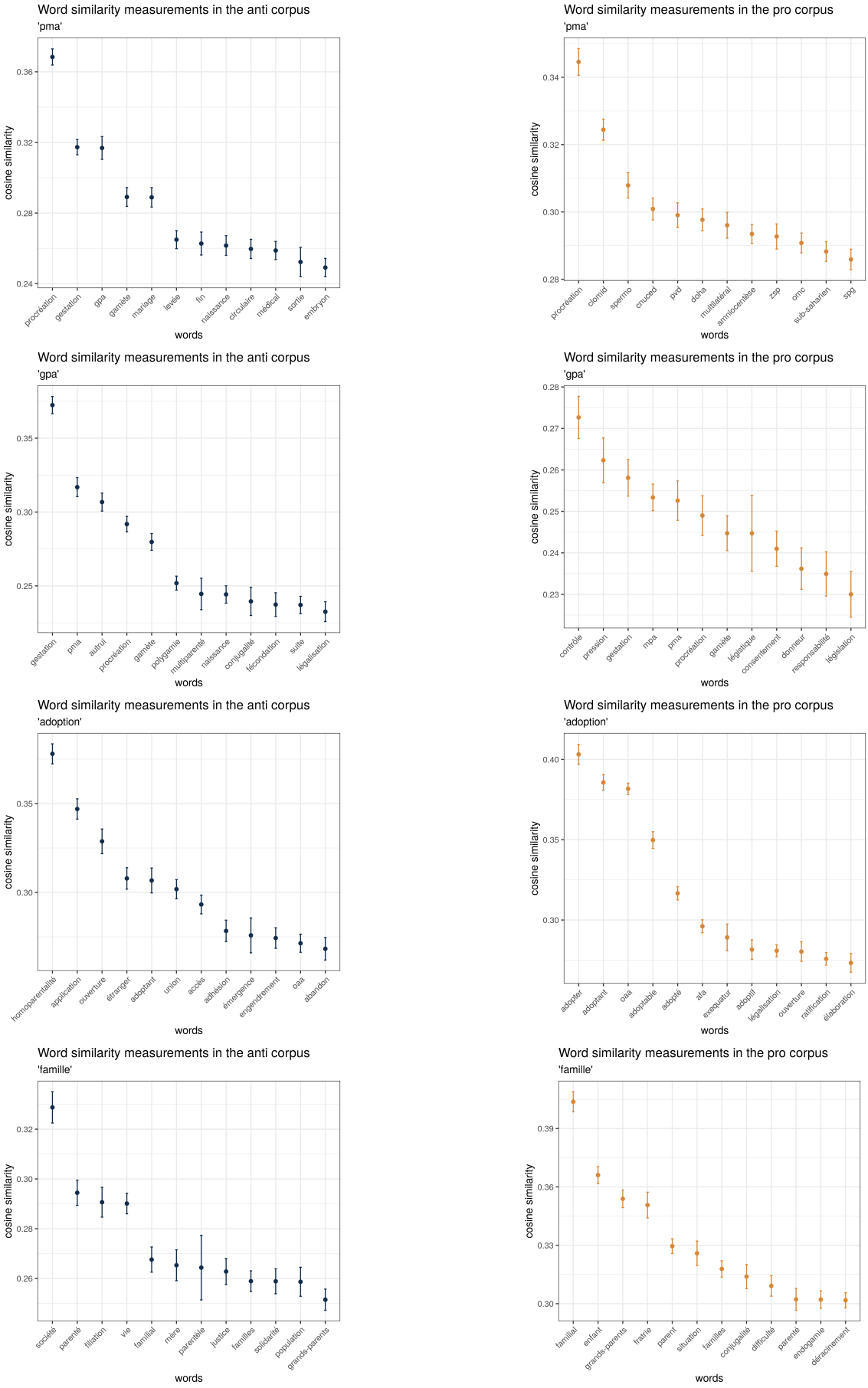

Figure 3: Here are more word comparisons among the test words. Results based on the *anti* models are on the left and those from the *pro* models are on the right. The figures are to be read the same as Figures 1 and 2.

A.2 Control words

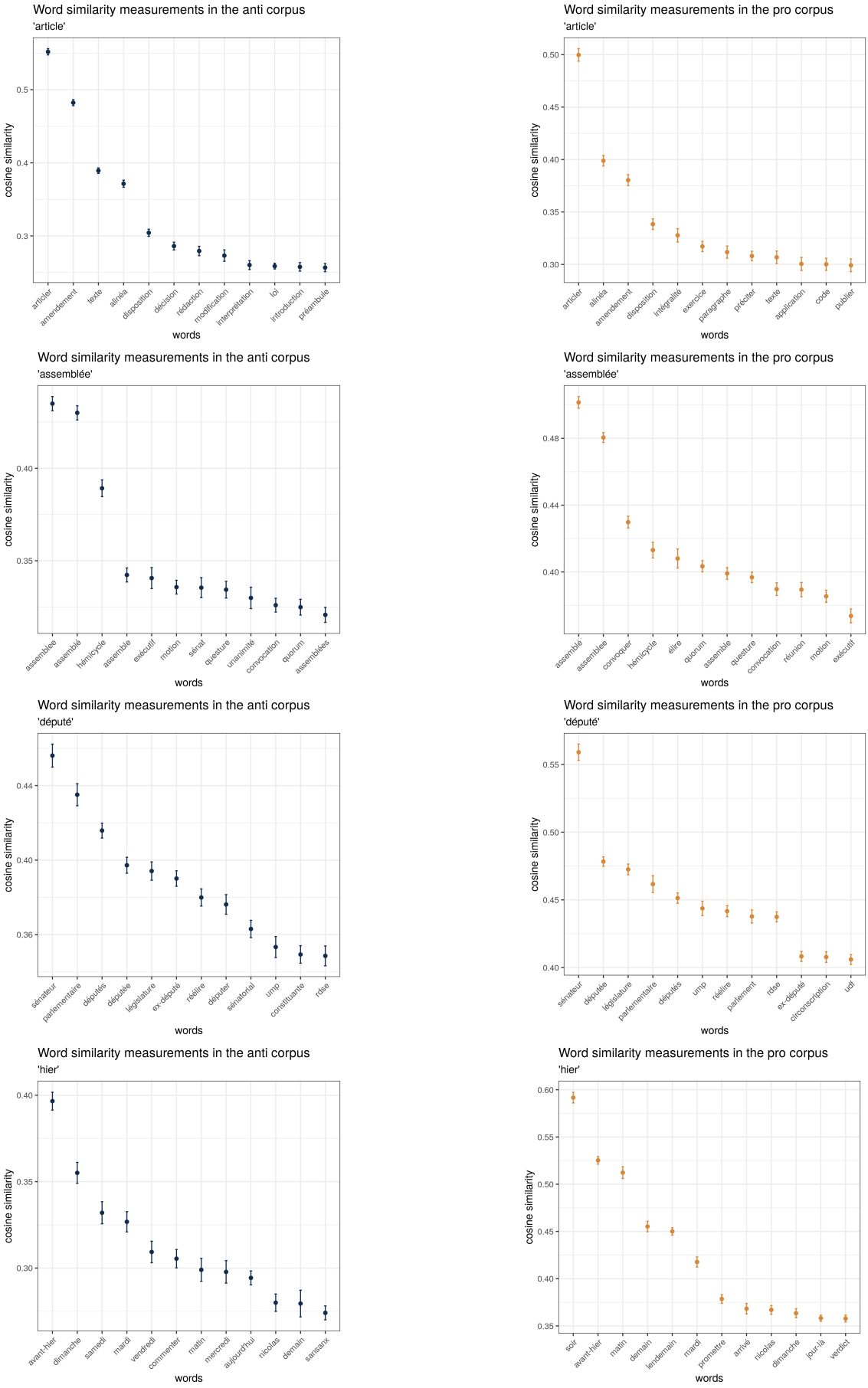

Figure 4: Here are more word comparisons among the control words. Results based on the *anti* models are on the left and those from the *pro* models are on the right. The figures are to be read the same as Figures 1 and 2.